Analyse morphologique fine pour la recherche d'information biomédicale

Vincent Claveau, Ewa Kijak

Analyse morphologique fine pour la recherche d'information biomédicale

Vincent Claveau[1,2] **Ewa Kijak**[1,3]

[1]*IRISA* – [2]*CNRS* – [3]*Univ. Rennes 1*
Campus de Beaulieu
F-35042 Rennes cedex
{vincent.claveau,ewa.kijak}@irisa.fr

RÉSUMÉ. *Dans le domaine biomédical, l'emploi de termes spécialisés est la clef de l'accès à l'information. Mais dans la plupart des langues indo-européennes, ces termes sont des constructions morphologiques complexes. Dans cet article, nous cherchons à identifier les différents éléments de sens composant ces termes et utilisons ces analyses pour améliorer la recherche d'information biomédicale. Nous présentons en particulier une approche automatique combinant alignement avec une langue pivot et apprentissage analogique permettant une analyse morphologique fine des termes. Ces analyses morphologiques sont utilisées pour améliorer l'indexation des documents médicaux. Les expériences rapportées dans cet article montrent le bien-fondé de cette approche avec des améliorations de MAP de +10% par rapport à un système de RI standard.*

ABSTRACT. *In the biomedical field, the key to access information is the use of specialized terms. However, in most of Indo-European languages, these terms are complex morphological structures. The aim of the presented work is to identify the various meaningful components of these terms and use this analysis to improve biomedical information retrieval. We present an approach combining an automatic alignment using a pivot language, and an analogical learning that allows an accurate morphological analysis of terms. These morphological analysis are used to improve the indexing of medical documents. The experiments reported in this paper show the validity of this approach through improvements in MAP of 10 % compared to a standard IR system.*

MOTS-CLÉS : *Morphologie, domaine biomédical, alignement, apprentissage analogique, indexation morpho-sémantique.*

KEYWORDS: *Morphology, biomedical field, alignment, analogical learning, morpho-semantic indexing.*

1. Introduction

Dans le domaine biomédical plus encore que dans d'autres domaines, l'emploi de termes spécialisés est la clef de l'accès à l'information. Ainsi, le thésaurus MeSH®(Medical Subject Headings - `www.nlm.nih.gov/mesh`) est développé pour indexer la célèbre base documentaire PubMED (`www.pubmed.gov`). Mais dans la plupart des langues indo-européennes, ces termes bio-médicaux se caractérisent par une grande complexité morphologique : ils sont le plus souvent issus de la composition dite néo-classique, de plusieurs racines, préfixes, et suffixes gréco-latins, et sont sujets à des variations. La maîtrise de cette complexité morphologique est donc essentielle pour la recherche d'information.

Dans cet article, nous nous intéressons à une technique d'analyse morphologique des termes bio-médicaux et à son utilisation dans une problématique de recherche d'information. Plus précisément, nous présentons une technique permettant de décomposer les termes en composants morphologiques, les morphes[1], et d'associer à ces derniers des informations sémantiques. Ces décompositions et les informations sémantiques associées sont ensuite exploitées pour permettre une indexation souple des documents dans un système de recherche d'information. Au contraire des travaux existants dans ce domaine (Deléger *et al.,* 2008, Markó *et al.,* 2005a), qui sont résolument basés sur l'expertise humaine, notre technique de décomposition est automatique et exploite les données existantes dans le domaine biomédical.

L'idée au cœur de notre technique de décomposition est d'exploiter l'aspect multilingue des bases terminologiques qui existent dans le domaine bio-médical. Nous utilisons le japonais comme langue pivot, et plus précisément les termes écrits en kanjis (charactères chinois utilisés en japonais), pour mener la décomposition des termes dans une autre langue à l'aide d'une technique d'alignement (Claveau *et al.,* 2011). En plus d'aider à la décomposition des termes, les kanjis servent aussi à étiqueter les morphes et fournissent ainsi une sorte de représentation sémantique. Ainsi, le terme anglais photochimiotherapy correspond au terme japonais par 光化学法 ; l'alignement de ces deux termes mène à la décomposition suivante : photo ↔ 光 ('lumière' en japonnais), chimio ↔ 化学 ('chimie'), therapie ↔ 法 ('therapie', 'traitement'). Comme cet exemple l'illustre, chaque morphe est associé à des kanjis qui peuvent servir de descripteurs plus souples que les termes eux-mêmes pour indexer des documents. Nous montrons en particulier dans cet article comment les correspondances construites entre morphes et kanjis peuvent être exploitées de différentes façons pour améliorer les résultats d'un système de recherche d'information.

L'analyse morphologique et l'indexation des documents qu'elle permet dépendent donc de l'étape essentielle d'alignement entre morphes et kanjis. Celle-ci est mise en œuvre par une technique originale, adaptée aux données traitées, reposant sur un algorithme *Forward-Backward* et par de l'apprentissage analogique. Après une des-

1. Dans cet article, nous distinguons les morphes, signes linguistiques élémentaires (segments), des morphèmes, classes d'équivalences de morphes partageant un signifié identique et des signifiants proches (Mel'čuk, 2006).

cription des travaux connexes en section 2, nous présentons cette technique d'alignement initialement proposée dans (Claveau *et al.,* 2011) et ses résultats en terme de décomposition morphologique en sections 3 et 4. L'utilisation des produits de ces décompositions dans un contexte de recherche d'information est détaillée en section 5. Les résultats obtenus sur une collection biomédicale sont présentés dans la section 6.

2. Travaux connexes

La morphologie a été étudiée dans de nombreux cadres (lexicologie, terminologie, recherche d'information...). Comme nous l'avons expliqué précédemment, elle tient un rôle particulier dans le domaine biomédical où les termes sont issus d'opérations morphologiques complexes mais régulières et productives. Malheureusement, au contraire des terminologies, très largement développées dans le domaine biomédical, il n'existe pas de bases de données de morphes enrichis avec des informations sémantiques, complètes et maintenues. Par ailleurs, la décomposition d'un terme en morphes, qui permettrait de tirer parti d'une telle base, reste aussi un problème. Notons enfin que les outils communément employés en RI pour tenir compte de la morphologie, notamment la racinisation, sont inadaptés à la complexité des constructions morphologiques à traiter.

En ce qui concerne l'utilisation de la morphologie comme outil d'analyse de termes ou de mots, il convient de distinguer deux visions du problème. Dans la vision lexématique, des relations sémantiques entre mots sont détectées en s'appuyant sur leur forme, mais sans besoin de décomposition (Grabar *et al.,* 2002, Claveau *et al.,* 2005, par exemple). À l'opposé de cet emploi implicite de la morphologie, la vision morphémique repose sur une décomposition en morphe en préalable à tout traitement. Beaucoup de travaux se placent dans ce cadre, qui est aussi celui adopté dans nos travaux. Les approches existantes sont soit fortement manuelles (Deléger *et al.,* 2008, Markó *et al.,* 2005a), ou plus automatiques. Dans ce dernier cas, les morphes sont souvent détectés comme des séquences de lettres répétées dans les mots d'un lexique (Kurimo *et al.,* 2010). Mais de telles techniques n'associent aucune information sémantique aux morphes détectés. À notre connaissance, aucune étude avant notre approche n'a exploré l'utilisation d'une langue pivot pour mener l'analyse morphologique. Cela s'explique par trois particularités du domaine que nous traitons : la morphologie biomédicale est connu pour être très régulière, la composition y est très utilisée, on dispose de ressources mutilingues.

D'un point de vue technique, l'utilisation d'une terminologie bilingue est à rapprocher des travaux en translittération, notamment du Katakan ou de l'Arabe (Tsuji *et al.,* 2002, Knight *et al.,* 1998, par exemple), ou même de traduction et de phonétisation. Le passage par la phonétique, souvent utilisé dans ce type de travaux, n'a pas de sens dans notre cas, de par la nature de l'écriture par kanji et par le fait que les termes en japonais et ceux en français ou anglais n'ont aucune raison d'être phonétiquement proches. Dans ce cadre, citons aussi les travaux de (Morin *et al.,* 2010) qui proposent de mettre en correspondance des termes en kanjis avec des termes en français en uti-

lisant des règles s'appuyant sur des indices morphologiques. Cependant, dans ce cas encore, les règles doivent être constituées à la main par un expert. De plus, ce travail ne s'intéresse qu'à un cas bien précis d'opération morphologique issue de la dérivation, et n'est pas adapté à traiter de la composition, l'opération morphologique régissant les composés néo-classiques. Rappelons enfin que des méthodes de traduction de termes biomédicaux considérant les termes comme de simples séquences de lettres ont été proposées (Claveau, 2009). Même si le but est différent, ces méthodes partagent logiquement certaines similarités avec l'approche présentée dans cet article. En effet, elles reposent sur des alignements des termes au niveau des lettres. Cela est effectué le plus souvent avec des algorithmes d'alignement 1-1, c'est-à-dire uniquement capables d'aligner une lettre (ou un caractère vide) du terme de la langue source avec un caractère du terme de la langue cible. Cependant, d'autres travaux récents sur la phonétisation ont souligné l'intérêt d'utiliser des algorithmes d'alignement *many-to-many* (Jiampojamarn *et al.,* 2007). C'est ce type d'algorithme qui sert de base à notre système de décomposition qui est présenté dans la section suivante.

Enfin, en ce qui concerne les traitements morphologiques en recherche d'information, la littérature est très riche. Le lecteur intéressé peut se reporter à (Moreau *et al.,* 2005) pour un panorama très complet. Si les résultats constatés dans ces études dépendent de nombreux facteurs (langue, outil morphologique, taille de la collection, domaine...), un consensus se dégage pour ce qui est des traitements simples comme la racinisation (*stemming*). Les outils de racinisation, simples et disponibles, permettent en effet d'améliorer les résultats d'un système de RI dans la plupart des cas. La lemmatisation, plus rare en RI, montre aussi de bons résultats. Il est important de noter que les seuls phénomènes morphologiques pris en compte par ces outils sont la flexion et la dérivation. La composition leur reste inaccessible puisqu'ils travaillent principalement sur les suffixes des mots. Plus récemment, les techniques d'analyse morphologique développées dans le cadre de MorphoChallenge ont été appliquées à des problèmes de RI (Kurimo *et al.,* 2009). Les auteurs ont là-aussi constaté un gain pour quelques langues, notamment le finnois, très compositionnel, mais les résultats sur l'anglais sont largement moins bons qu'en utilisant un simple *stemming* de Porter.

3. Décomposition par alignement

Comme nous l'avons expliqué, notre technique de décomposition repose sur l'alignement avec des termes d'une langue pivot. Cette approche fait donc l'hypothèse que les termes en kanjis ont des constructions parallèles à celles des termes dans la langue étudiée. Cette hypothèse peut paraître forte, mais les résultats donnés ci-après montrent que cette hypothèse est raisonnable dans la plupart des cas. Il est important de noter que l'utilisation des kanjis n'est pas fortuite. Ils n'ont pas de morpholgie qui s'y appliquent, leur forme est donc invariable quelle que soit leur position dans le terme. De ce fait, envisager leur segmentation revient à examiner peu de combinaisons, contrairement à un terme en alphabet latin contenant beaucoup de lettres. Ils sont indépendants des racines gréco-latines employées en français ou en anglais, ce

qui empêche d'apprendre des régularités qui seraient fortuites. Enfin, toujours grâce à l'absence de morphologie, leur sens est directement exploitable à l'aide de dictionnaires, contrairement à des morphèmes. C'est différents point en font un excellent langage pivots par rapport à d'autres alternatives comme l'allemand.

La technique d'alignement que nous utilisons repose sur un algorithme *Expectation-Maximization* (EM) (Jiampojamarn *et al.,* 2007, pour un exemple d'utilisation), rappelé dans la sous-section suivante. La seconde sous-section présente la modification apportée à cet algorithme pour prendre en compte au mieux les spécificités morphologiques de nos données (Claveau *et al.,* 2011).

3.1. *Alignement EM*

L'algorithme d'alignement est relativement standard : il s'agit d'un algorithme *Baum-Welch* étendu pour pouvoir gérer des sous-séquences de symboles et non seulement des alignements 1-1. Les longueurs maximales des sous-séquences dans la langue 1 et dans la langue 2 sont données en paramètres et notées $maxX$ et $maxY$ ci-après. Dans notre cas, l'algorithme prend en entrée des termes dans la langue étudiée (anglais ou français) avec leur traduction en kanji. Ces paires sont issues de bases terminologiques multilingues, en particulier de l'UMLS.

Pour chaque paire de termes (x^T, y^V) à aligner (T et V sont les longueurs des termes en caractères/kanjis), l'algorithme EM (algorithme 1) procède de la manière suivante. La phase d'*Expectation* recense les comptes partiels de tous les alignements possibles entre sous-séquences de kanjis et de caractères. Ces comptes sont conservés dans la table γ ; ils sont ensuite utilisés dans la phase de *Maximization* pour estimer les probabilités d'alignement de la table δ.

La phase d'*Expectation* implémente une approche *forward-backward* (algorithme 2) : elle estime les probabilités *forward* notées α et *backward* notées β. À chaque position t, v d'une paire de termes, $\alpha_{t,v}$ est la somme des probabilités de tous les alignements possibles du début des termes jusqu'à ces positions (x_1^t, y_1^v), calculés à partir des probabilités d'alignement courantes contenues dans δ (voir algorithme 4). De manière similaire, $\beta_{t,v}$ est calculé en considérant la fin des termes (x_t^T, y_v^V). Ces probabilités α et β sont ensuite utilisées pour re-estimer les comptes de γ. Dans sa version originale, la phase de *Maximization* (algorithme 3) consiste simplement à calculer les probabilités δ en normalisant les comptes dans γ.

Algorithm 1 *Algorithme EM*

Input : liste de paires (x^T, y^V), $maxX$, $maxY$
while δ est modifié **do**
 initialisation de γ à 0
 for all paire (x^T, y^V) **do**
  γ = Expectation(x^T, y^V, $maxX$, $maxY$, γ)
 δ = Maximization(γ)
return δ

Algorithm 2 *Expectation*

Input : (x^T, y^V) , $maxX$, $maxY$, γ
$\alpha :=$ Forward-many2many(x^T , y^V , $maxX$, $maxY$)
$\beta :=$ Backward-many2many(x^T , y^V , $maxX$, $maxY$)
if $\alpha_{T,V} > 0$ **then**
 for $t = 1...T$ **do**
  for $v = 1...V$ **do**
   for $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
    for $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
     $\gamma(x^t_{t-i+1}, y^v_{v-j+1})$ += $\frac{\alpha_{t-i,v-j}\delta(x^t_{t-i+1}, y^v_{v-j+1})\beta_{t,v}}{\alpha^{T,V}}$
return γ

Algorithm 3 *Maximization*

Input : γ
for all sous-séquence a t.q. $\gamma(a, \cdot) > 0$ **do**
 for all sous-séquence b t.q. $\gamma(a, b) > 0$ **do**
  $\delta(a, b) = \frac{\gamma(a,b)}{\Sigma_x \gamma(a,x)}$
return δ

Algorithm 4 *Forward-many2many*

Input : (x^T , y^V) , $maxX$, $maxY$
$\alpha_{0,0} := 1$
for $t = 0...T$ **do**
 for $v = 0...V$ **do**
  if $(t > 0 \vee v > 0)$ **then**
   $\alpha_{t,v} = 0$
  if $(v > 0 \wedge t > 0)$ **then**
   for $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
    for $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
     $\alpha_{t,v}$ += $\delta(x^t_{t-i+1}, y^v_{v-j+1})\alpha_{t-i,v-j}$
return α

Ce processus EM est répété tant que les probabilités dans δ changent. Une fois la convergence atteinte, les alignements sont finalement produits comme ceux maximisant $\alpha(T, V)$. En plus de ces alignements, nous conservons également les probabilités d'alignement des sous-mots collectées dans δ, qui nous sont utiles pour des traitements de certains termes en RI (cf. section 5.2).

Cette technique est assez proche de celle utilisée en traduction artificielle mais quelques différences peuvent être soulignées. Cette approche ne permet pas de gérer la distorsion, c'est-à-dire le ré-ordonnancement de morphes. Ces techniques intégrant la distortion peuvent sembler plus adaptée puisqu'elle permettraient de lever notre hypothèse sur l'ordonnancement similaire des morphes. Cependant, elles nécessitent de

ce fait beaucoup plus de données d'entraînement sans assurer de meilleurs résultats puisque la tâche d'estimation des paramètres du modèle d'alignement se trouve complexifiée (voir à ce titre les résultats de GIZA++ présentés ci-après). En revanche, la gestion de la fertilité, c'est-à-dire la possibilité d'avoir des sous-chaînes vides, n'est pas présentée ici par souci de place mais peut être prise en compte simplement avec cet algorithme.

3.2. *Normalisation morphologique automatique*

La phase de *Maximization* calcule simplement les probabilités de traduire une sous-chaîne de kanjis en une sous-séquence de lettres. Les particularités de nos données, et plus précisément la variation morphologique, y est mal prise en compte, ce qui conduit à dégrader les résultats. Par exemple, pour le kanji 菌 ('*bacteria*'), la table δ peut recenser plusieurs traductions : bactérie, ou bien bactério (comme dans bactério/lyse), ou encore bactéri (dans myco/bactéri/ose), chacune avec une certaine probabilité. Cette dispersion des probabilités pour un même morphème est préjudiciable. L'adaptation que nous avons proposée a pour but de rendre la phase de *Maximization* capable de gérer ces différentes variantes, et donc de grouper les différents morphes en un unique morphème. Pour ce faire, nous utilisons une technique simple reposant sur le raisonnement analogique.

3.2.1. *Analogie*

Une analogie est une relation entre 4 éléments que nous notons $a : b :: c : d$ et qui peut se lire a *est à* b *ce que* c *est à* d (Lepage, 2000, pour un panorama complet). Les analogies ont été utilisées dans beaucoup de problèmes de traitement automatique des langues, notamment la traduction de phrases (Lepage, 2000) ou de termes (Langlais *et al.,* 2007, Langlais *et al.,* 2008), ou la structuration de terminologie (Claveau *et al.,* 2005). Nous nous appuyons sur ces derniers travaux pour formaliser notre problème de normalisation des morphes. Dans ce cadre, une analogie serait : dermato : dermo :: hémato : hémo. Si l'on sait en plus que dermato et dermo sont deux morphes d'un même morphème, on peut inférer que c'est aussi le cas pour hémato et hémo.

Comme pour la plupart des langues indo-européennes, les variations morphologiques simples du français ou de l'anglais se font principalement par préfixation et suffixation. Dans notre approche, ces analogies sont implémentées par des règles de réécriture s'appliquant aux préfixes et aux suffixes : la règle de réécriture permettant de passer de dermato à dermo permet aussi de passer de hémato à hémo. La base est la plus longue sous-chaîne commune entre les deux morphes (notée lcss, *longest common substring*). Dans l'exemple précédent, en notant $\oplus$ l'opérateur de concaténation, cette règle de réécriture r serait : $r = \text{lcss}(\text{morph}_1,\text{morph}_2) \ominus \text{ato} \oplus \text{o}$.

3.2.2. *Normalisation par analogie*

La méthode précédente est intéressante si l'on dispose d'exemples de morphes que l'on sait appartenir au même morphème (comme dermato et dermo). Nous ne disposons pas de tels exemples, mais il est possible d'utiliser une technique d'amorçage. Celle-ci consiste à considérer que deux morphes partageant une grande sous-chaîne commune et connus dans γ comme traductions très probables d'une même sous-chaîne de kanjis sont considérés comme des exemples. Ces amorces sont générées à chaque itération. À partir de ces amorces, les règles de réécriture de préfixation et suffixation sont construites et permettent de trouver d'autres morphes en analogie (au contraire des paires d'amorçage, celles-ci peuvent ne partager qu'une petite sous-chaîne commune). Plus une règle s'applique souvent pour les amorces, plus elle peut être considérée comme fiable. On conserve donc les règles les plus fiables à chaque itération. Ce processus est donc entièrement automatique.

Tous les morphes détectés en analogie avec les amorces sont considérés comme appartenant au même morphème. Il est donc maintenant possible d'estimer les probabilités δ en prenant en compte les différentes variantes des morphes. Cette nouvelle version de la *Maximization* assure que tous les morphes supposés appartenir au même morphème aient des probabilités égales et renforcées.

4. Evaluation de l'alignement

4.1. *Données et vérité terrain*

Les données utilisées dans nos expériences sont issues du MetaThesaurus de l'UMLS (Tuttle *et al.,* 1990). Le MetaThésaurus groupe plusieurs terminologies dans différents langages et associe à chaque terme un *identifieur conceptuel unique* (CUI). Les CUI sont indépendants des langues et permettent donc d'extraire facilement des listes de termes dans la langue souhaitée avec leurs équivalents en japonais. Dans notre cas, nous nous sommes intéressés au français et à l'anglais. Dans ces deux cas, nous ne considérons dans l'UMLS que les termes japonais écrit en kanjis et pour le français ou l'anglais, que les termes simples, c'est-à-dire composés d'un seul mot. Un marqueur de fin de terme (' ;') est ajouté à ces derniers pour distinguer les suffixes.

Ce sont finalement 14 000 paires de termes anglais-kanjis et 8 000 paires français-anglais qui sont constituées. Parmi ces paires, 1 600 paires pour le français et 500 pour l'anglais servent de vérité terrain pour évaluer notre approche. Elles ont été décomposées et alignées à la main et vont nous permettre d'évaluer les résultats de notre technique d'alignement.

4.2. *Résultats d'alignement*

Nous évaluons la performance de l'alignement en terme de précision : l'alignement d'une paire de termes est considérée correcte si tous ses composants sont correctement découpés et alignés (ce serait l'équivalent du *sentence error rate* en traduction).

Pour chaque paire, l'algorithme EM indique les probabilités de l'alignement proposé. Il nous est donc possible de ne considérer que les alignements dont la probabilité est supérieure à un certain seuil. En faisant varier ce seuil, on peut ainsi calculer une précision en fonction du nombre de termes alignés (nombre de termes dont le score est supérieur au seuil). Les figures 1 et 2 présentent respectivement les résultats obtenus sur les paires de test pour le français et l'anglais. L'influence de la normalisation morphémique par analogie est illustrée en faisant apparaître les résultats d'alignement avec et sans cette modification de l'algorithme. À des fins de comparaisons, nous rapportons aussi les résultats de GIZA++ (Och *et al.,* 2003), un algorithme d'alignement de référence dans le domaine de la traduction artificielle. Les différents modèles IBM et paramètres associés disponibles dans GIZA++ ont été testés ; les courbes affichées correspondent aux meilleurs résultats obtenus (IBM modèle 4 sans distorsion).

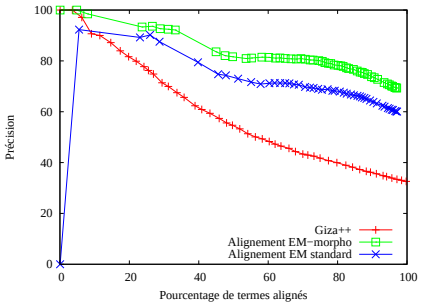

Figure 1. *Précision de l'alignement français-kanji selon le nombre de paires alignées*

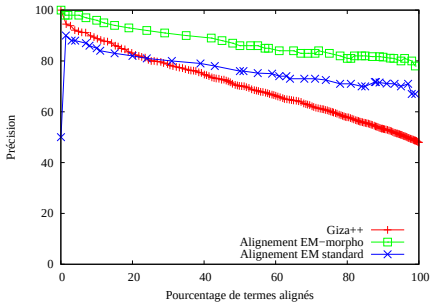

Figure 2. *Précision de l'alignement anglais-kanji selon le nombre de paires alignées*

On note dans ces figures les très bons scores de notre approche, avec dans le pire des cas (c'est-à-dire quand tous les alignements sont conservés), 70 % de précision pour le français et 80 % pour l'anglais. Comme attendu, l'intérêt de la normalisation morphémique apparaît clairement avec un gain constant d'environ 10 %. La normalisation a aussi un autre intérêt calculatoire puisqu'elle réduit le nombre nécessaire d'itérations de l'algorithme EM et permet donc une convergence plus rapide.

Un examen manuel des résultats montre sans surprise que la plupart des erreurs sont causées par la mise en défaut de notre hypothèse de départ : certaines paires ne se décomposent pas de la même façon en kanjis et dans l'autre langue considérée. Par exemple, le terme français anxiolytiques se traduit en japonais par une suite de kanji signifiant littéralement 'médicament pour la dépression'. Certaines erreurs sont aussi causées par le fait qu'au moins un des deux termes n'est pas un composé

néo-classique, comme par exemple méninges (alors que sa traduction est bien une composition : 膜 signifie 'membrane du cerveau'). D'autres erreurs sont causées par un manque de données d'entraînement : certains morphes ou séquence de kanjis n'apparaissent qu'une fois dans les données d'entraînement, ou bien toujours combinés avec les mêmes autres morphes, ce qui rend les comptes effectués par l'algorithme peu fiables.

5. Analyse morpho-sémantique pour la recherche d'information

Comme nous l'avons dit précédemment, la recherche d'information biomédicale a des caractéristiques particulières dues à l'utilisation des termes spécialisés. À ce titre, le bien-fondé de prendre en compte des informations morphologiques riches sur ces termes a déjà été montré, mais uniquement en utilisant des ressources développées manuellement (Markó *et al.,* 2005b). Dans cette section, nous explorons les différentes utilisations de notre technique de décomposition automatique dans un cadre de RI, sur des documents biomédicaux en anglais. Nous présentons tout d'abord les différentes informations qu'il est possible d'extraire des analyses morphologiques produites par alignement, puis nous indiquons la mise en œuvre adoptée pour inclure ces informations au sein d'un système de RI.

5.1. *Graphes morpho-sémantiques*

Une fois l'alignement effectué, il est possible d'étudier les correspondances récurrentes entre morphes et kanjis dans les données finalement alignées. Plus un morphe est aligné souvent avec une séquence de kanjis, plus le lien sémantique entre eux est sûr. Tous ces liens peuvent être utilisés pour construire un graphe dont les nœuds sont les kanjis ou les morphes ou même les morphèmes (morphes groupés par analogie lors de la phase de maximisation), et les liens représentent donc les correspondances trouvées, pondérés par leur nombre. La figure 3 montre un exemple jouet d'un tel graphe anglais-kanji. La taille des arcs est proportionnelle à la force du lien.

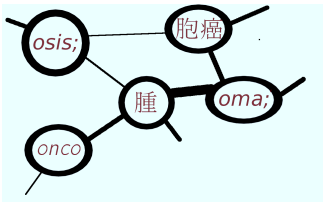

Figure 3. *Graphe morphème-kanji*

Cette représentation nous permet de mettre en lumière les différents types de relations entre morphèmes. Cela se fait simplement en explorant les voisinages des mor-

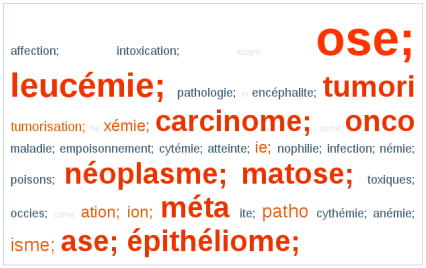

Figure 4. *Nuage d'affinités du 1er ordre du suffixe* ome

Figure 5. *Nuage d'affinités du 2nd ordre du morphème* gastro

phèmes : chaque noeud reçoit une certaine quantité d'énergie qu'il propage, proportionnellement au poids des arcs, à ses voisins ; on peut ainsi lister les nœuds ayant été atteints, et la quantité d'énergie reçue illustre la proximité avec le nœud de départ. Par exemple, la figure 4 montre, sous forme de nuages de tags, les morphèmes les plus proches du morphème ome, un suffixe de termes liés au cancer. La taille et la couleur des nœuds illustrent la proximité. Ces nœuds proches sont supposés conceptuellement liés et doivent être des synonymes ou des quasi-synonymes de ome. Il est intéressant de voir que non seulement des suffixes sont trouvés mais aussi des préfixes comme onco.

L'alignement et la segmentation produits par notre algorithme rendent également possible l'étude des cooccurrences des morphèmes anglais (ou français) entre eux. Il est par exemple possible d'étudier les affinités de premier ordre, c'est-à-dire quels sont les morphèmes fréquemment associés ensemble. Plus intéressant, on peut également étudier les affinités de second ordre, c'est-à-dire les morphèmes partageant les mêmes morphèmes cooccurrents (mêmes contextes). Les affinités de second-ordre doivent nous permettre de grouper les morphèmes par paradigme. Par exemple, le nuage de mots en figure 5 montre les morphèmes associés avec gastro (estomac) selon l'affinité de second ordre. On constate comme attendu que ces voisins identifient pour la plupart des organes, et les plus proches désignent des organes les plus proches de l'estomac.

Ces informations de différentes natures permettent d'identifier des relations entre morphèmes et donc entre termes. L'évaluation directe des groupements produits n'est pas possible faute de référence, mais ce sont ces informations qu'on se propose d'utiliser dans un système de RI, offrant ainsi une évaluation indirecte de la pertinence des analyses morphologiques produites.

5.2. *Représentation morphémique pour la RI*

Pour intégrer les informations morphologiques dans le système de RI, nous adoptons une représentation simple : les documents et les requêtes sont considérés comme des sacs de morphèmes et de mots. Les morphèmes sont ceux obtenus par décomposition des termes biomédicaux ou ceux obtenus par affinités de second ordre avec ces premiers. Le but est bien sûr de pouvoir apparier une requête contenant un terme comme stomachalgia avec un document contenant gastrodynia.

Lors de l'indexation de la collection, les termes sont donc décomposés. Deux cas peuvent se présenter : soit le terme est un terme apparaissant dans les paires d'alignement, soit non. Dans le premier cas, nous récupérons sa décomposition telle que produite par l'alignement. Dans le second cas, nous exploitons les probabilités collectées dans la table δ pour générer la traduction la plus probable. Pour ce faire, nous utilisons une approche très simple : les probabilités de traductions des morphes dans δ sont utilisées dans un algorithme de Viterbi pour générer la traduction en kanji de probabilité maximale. Nous n'utilisons pas de modèle de langue. Il est important de noter que cette traduction produit en même temps la décomposition voulue du terme initial en assignant à chacun des morphes sa traduction en kanjis. Ce processus de traduction correspond donc bien à l'analyse morpho-sémantique d'un terme inconnu, c'est-à-dire absent des paires utilisées pour l'alignement. Dans chacun de ces deux cas, nous utilisons aussi un autre produit de l'alignement : il s'agit des règles de réécriture. Collectées à la dernière itération de l'algorithme EM, celles-ci permettent de détecter les morphes appartenant au même morphème. Elles nous permettent de mettre en correspondance une requête contenant hemo avec un document contenant haemo, hemato ou encore emia ;.

Une *baseline* et quatre systèmes d'indexation utilisant ces décompositions ont été testés. Tous reposent sur un modèle vectoriel utilisant la pondération BM-25 d'Okapi (Robertson *et al.,* 1998) avec les valeurs standard pour les paramètres b, k_1, k_3. La baseline fait une indexation classique des documents avec une racinisation de Porter.
1 – Le premier est basé morphème ; il considère simplement les morphèmes issus de la décomposition des termes rencontrés dans les documents et les requêtes comme des mots à indexer. La pondération du morphème tient compte de la probabilité de décomposition ; elle est définie comme le produit de cette probabilité avec le poids du morphème tel que fourni par le modèle de RI utilisé.
2 – Le second est basé kanjis ; les termes sont là-aussi décomposés, mais ce sont les kanjis proches qui sont utilisés comme descripteurs. Ces kanjis proches sont ceux identifiés dans le voisinage des morphes issus de la décomposition.
3 – Le troisième système reprend la représentation en morphème du premier système mais étend les requêtes avec les affinités de premier ordre de leurs morphèmes. Les extensions sont pondérées selon leur proximité dans le graphe et par le poids du morphème qu'elles étendent.
4 – Le dernier système est identique au troisième mais utilise les affinités de second-ordre pour étendre les requêtes.

	baseline (BM-25 + stemming)	Système basé morphème	Système basé kanji
MAP	29.93	33.94 (+13.4 %)	32.76 (+9.5 %)
IAP	31.74	35.55 (+12 %)	34.49 (+8.6 %)
R-prec	35.28	39.64 (+12.3 %)	38.59 (+9.4 %)
P@5	69.87	73.45 (+5.1 %)	71.70 (+2.6 %)
P@10	67.99	71.31 (+4.9 %)	*69.65 (+2.4 %)*
P@50	52.98	56.90 (+7.4 %)	55.24 (+4.3 %)
P@100	40.86	44.56 (+9.1 %)	43.39 (+6.2 %)
P@500	15.11	17.21 (+13.9 %)	16.92 (+12 %)
P@1000	8.72	10.10 (+15.86 %)	9.95 (+14.2 %)

Tableau 1. *Performances des différents systèmes de RI sur la collection OHSUMED, avec les requêtes TREC.*

6. Expériences de RI biomédicale

6.1. *Contexte expérimental*

Pour les expériences rapportées ci-après, nous utilisons le jeu de données construit pour la tâche de filtrage de la conférence TREC-9. Ce jeu de donnée s'appuie lui-même sur la collection de document OHSUMED, qui est composée de 350 000 résumés d'articles scientifiques extraits de MEDLINE. Dans TREC-9, 4 000 requêtes de filtrage et leur jugement de pertinence ont été développés. Ces requêtes sont composées de plusieurs champs : le sujet, qui est un terme MeSH, et une définition de ce terme. Bien que développé initialement pour le filtrage, nous utilisons ce jeu de donnée comme une collection standard de RI, et ne considérons que le champ sujet pour former nos requêtes.

6.2. *Résultats*

La table 1 présente les résultats du système *baseline* et des deux premiers systèmes basés sur la représentation en morphèmes et en kanjis. Les performances des différents systèmes y sont évaluées à l'aide des mesures usuelles : précision sur les 5, 10,... 1000 premiers documents (P@x), MAP, précision moyenne interpolée (IAP) et R-précision.Pour vérifier la significativité des différences constatées entre les systèmes, nous effectuons un test statistique de Wilcoxon ($p = 0.05$) (Hull, 1993) ; les différences avec le système *baseline* jugées non statistiquement significatives sont en italiques.

Le sytème basé morphème, reposant donc simplement sur la décomposition des termes et le regroupement en morphème, obtient de très bons résultats avec un gain en MAP de 13 %. Comme attendu, la décomposition améliore plus particulièrement

les performances en fin de liste (P@100 et supérieurs) puisqu'elle permet de ramener des documents même s'ils ne contiennent pas les termes de la requêtes. Le système basé kanji obtient des performances assez semblables. Le gain espéré de passer à une représentation plus générique que les morphèmes n'est pas réalisé. Il semble que cette représentation soit trop générique pour certaines requêtes et apporte peu d'information supplémentaire par rapport au morphème pour d'autres. Par ailleurs, aucune sélection n'est faite sur les morphèmes à traduire ou non, et certains kanjis trouvés en traduction ont des propriétés (fréquence documentaire) différentes du morphème de départ puisqu'ils peuvent se trouver comme traduction de différents morphèmes. Une technique de pondération tenant compte des fréquences documentaires initiales semble une perspective importante pour développer ce type de système.

Le tableau 2 illustre les résultats des deux derniers systèmes proposés reposant sur l'extension de requêtes. Les systèmes à base d'extension ont des résultats plus contras-

	baseline (BM-25 + stemming)	Système avec affinités 1er ordre	Système avec affinités 2nd ordre
MAP	29.93	34.40 (+14.9 %)	28.74 (-3.9 %)
IAP	31.74	36.63 (+15.4 %)	30.80 (-2.9 %)
R-prec	35.28	39.92 (+13.2 %)	34.38 (-2.6 %)
P@5	69.87	71.76 (+2.7 %)	*68.65 (-1.7 %)*
P@10	67.99	70.46 (+3.6 %)	*66.20 (-2.6 %)*
P@50	52.98	56.30 (+6.7 %)	50.50 (-4.68 %)
P@100	40.86	44.69 (+9.4 %)	39.07 (-4.38 %)
P@500	15.11	17.98 (+18.9 %)	*15.01 (-0.64 %)*
P@1000	8.72	10.56 (+21.1 %)	*8.77 +0.66 %)*

Tableau 2. *Performances des différents systèmes de RI sur la collection OHSUMED, avec les requêtes TREC.*

tés. L'extension à l'aide d'affinités du premier ordre donne de très bons résultats, avec une précision un peu dégradée en début de liste mais un rappel amélioré. En revanche, les affinités de second ordre produisent des résultats largement moins bons qu'avec la décomposition morphologique seule. Il semble que ces affinités soient trop éloignées sémantiquement et ramènent trop de documents jugés non pertinents.

7. Conclusion et perspectives

Bien que la morphologie soit étudiée de longue date en RI, et malgré la disponibilité d'outils simples comme les *stemmers*, la morphologie reste un enjeu d'importance, au même titre que d'autres phénomènes linguistiques pouvant sembler plus complexes à modéliser. Comme nous l'avons vu, les outils actuels ne sont pas suffisants lorsqu'ils sont confrontés à des opérations morphologiques compliquées comme celles ayant cours dans le domaine médical. À ce titre, nos résultats s'inscrivent dans la continuité de ceux rapportés par d'autres sur l'intérêt de prendre en compte ces

phénomènes (Markó *et al.,* 2005a, Deléger *et al.,* 2008), mais à notre connaissance, ce sont les premiers à proposer un processus entièrement automatique, sans intervention humaine, et directement applicable à de nombreuses langues. Bien sûr, il repose sur la disponibilité de terminologies multilingues, mais celles-ci, au contraire de bases de connaissances morphologiques, sont largement disponibles.

Ce travail ouvre de nombreuses perspectives. Tout d'abord d'un point de vue technique, il serait intéressant de produire des décompositions de termes non plus linéaires, mais hiérarchisées (par exemple, gastroentérite serait analysé en [gastro | entér] ite). Nous pensons que ces décompositions nous permettraient dans un cadre de RI de pondérer plus efficacement les morphes et de choisir plus facilement ceux pouvant être étendus par des morphes liés sémantiquement. Pour ce faire, on peut là encore imaginer exploiter les kanjis dont certains ont une fonction syntaxique connue (certains sont notamment des prédicats attendant un agent ou un objet). Outre ces considérations syntaxiques au sein des termes, il est aussi possible d'exploiter les liens sémantiques entre kanjis, facilement récupérables à partir de dictionnaires généralistes japonais, pour aider à établir les liens sémantiques entre morphes.

Enfin, le domaine biomédical est aussi très riche en termes complexes (composés de plusieurs mots). Une adaptation de notre approche d'analyse morphologique pouvant s'appliquer à ces termes est à l'étude. La difficulté est alors de gérer les différents ordonnancements des mots composant le terme, et donc d'autoriser la distorsion. Pour la RI, l'enjeu est cependant important puisque ces termes sont connus pour les nombreuses variations qu'ils peuvent subir, ce qui empêche la mise en correspondance des documents et des requêtes contenant des variantes différentes d'un même terme.

8. Bibliographie

Claveau V., « Translation of Biomedical Terms by Inferring Rewriting Rules », *in* , V. Prince, , M. Roche (eds), *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, IGI - Global, 2009.

Claveau V., Kijak E., « Morphological Analysis of Biomedical Terminology with Analogy-Based Alignment », *Proceedings of the RANLP conference*, Hissar, Bulgarie, 2011.

Claveau V., L'Homme M.-C., « Structuring Terminology by Analogy-Based Machine Learning », *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhaguen, Denmark, 2005.

Deléger L., Namer F., Zweigenbaum P., « Morphosemantic parsing of medical compound words : Transferring a French analyzer to English. », *International Journal of Medical Informatics*, vol. 78, n° Supplement 1, p. 48-55, 2008.

Grabar N., Zweigenbaum P., « Lexically-based terminology structuring : Some inherent limits », *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan, 2002.

Hull D., « Using Statistical Testing in the Evaluation of Retrieval Experiments », *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis, 1993.

Jiampojamarn S., Kondrak G., Sherif T., « Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion », *Proc. of the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, 2007.

Knight K., Graehl J., « Machine Transliteration », *Computational Linguistics*, vol. 24, n° 4, p. 599-612, 1998.

Kurimo M., Creutz M., Turunen V., « Morpho challenge evaluation by information retrieval experiments », *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, Springer-Verlag, Berlin, Heidelberg, p. 991-998, 2009.

Kurimo M., Virpioja S., Turunen V. T., *(Eds), Proceedings of the MorphoChallenge 2010*, Espoo, Finlande, 2010.

Langlais P., Patry A., « Translating Unknown Words by Analogical Learning », *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, p. 877-886, June, 2007.

Langlais P., Yvon F., Zweigenbaum P., « Translating Medical Words by Analogy », *Proc. of the workshop on Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) 2008*, Washington, DC, 2008.

Lepage Y., « Languages of analogical strings », *Proc. of the 18th conference on Computational linguistics, COLING'00*, Universität des Saarlandes, Saarbrücken, Germany, 2000.

Markó K., Schulz S., Han U., « Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain », *Methods of Information in Medicine*, 2005a.

Markó K., Schulz S., Medelyan O., Hahn U., « Bootstrapping Dictionaries for Cross-Language Information Retrieval », *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brésil, 2005b.

Mel'čuk I., *Aspects of the Theory of Morphology*, Trends in Linguistics. Studies and Monographs, Mouton de Gruyter, Berlin, March, 2006.

Moreau F., Sébillot P., Contributions des techniques du traitement automatique des langues à la recherche d'information, Technical Report n° 1690, IRISA, 2005.

Morin E., Daille B., « Compositionality and lexical alignment of multi-word terms », *Language Resources and Evaluation (LRE)*, 2010.

Och F. J., Ney H., « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, vol. 29, n° 1, p. 19-51, 2003.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199-210, 1998.

Tsuji K., Daille B., Kageura K., « Extracting French-Japanese Word Pairs from Bilingual Corpora based on Transliteration Rules », *Proc. of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Spain, 2002.

Tuttle M., Sherertz D., Olson N., Erlbaum M., Sperzel D., Fuller L., Neslon S., « Using Meta-1 – the 1st Version of the UMLS Metathesaurus », *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, Washington, USA, p. 131-135, 1990.